How to kickstart an AI venture without proprietary data: AI start-ups have a chicken and egg problem — here is how to solve it

Received (in revised form): 28th May, 2021

Kartik Hosanagar

Professor, Wharton School of the University of Pennsylvania, USA

Kartik Hosanagar is a professor at the Wharton School of the University of Pennsylvania. He is an entrepreneur and founder of Jumpcut Media and Yodle. Kartik is the author of *A Human's Guide to Machine Intelligence*, and faculty director of Wharton's AI for Business Initiative.

Twitter: @khosanagar; E-mail: kartikh@wharton.upenn.edu

Monisha Gulabani

Research Assistant, Wharton AI for Business, USA

Monisha Gulabani is a Research Assistant at Wharton AI for Business.

E-mail: monisha.gulabani@gmail.com

Abstract Even when entrepreneurs have innovative ideas for applying AI to real-world problems, they can encounter a unique challenge to kickstarting their AI ventures. Today's AI systems need to be trained on large datasets, which poses a chicken-and-egg problem for entrepreneurs. Established companies with a sizable customer base already have a stream of data from which they can train AI systems, build new products and enhance existing ones, generate additional data, and rinse and repeat. Entrepreneurs have not yet built their company, so they do not have data, which means they cannot create an AI product as easily; however, this challenge can be navigated with a strategic approach. This paper presents five strategies that can help entrepreneurs access the data they need to break into the AI space, as well as examples of how these strategies have been used by other companies, particularly in their early stages. Specifically, the paper discusses how entrepreneurs can: 1) start by offering a service that has value without AI and that generates data; 2) partner with a non-tech company that has a proprietary dataset; 3) crowdsource the (labelled) data they need; 4) make use of public data; and 5) rethink the need for data entirely and instead use expert systems or reinforcement learning to kickstart their AI ventures.

KEYWORDS: artificial intelligence (AI), machine learning (ML), data, entrepreneurship, innovation, AI start-ups, technology

INTRODUCTION

Banks lose billions of dollars to credit card fraud on an annual basis. Better detection or prediction of fraud would be incredibly valuable. As such, entrepreneurs might momentarily consider the possibility of convincing a bank to share their transactional data in the hope of building a better fraud detection algorithm. The catch, unsurprisingly, is that no major bank is willing to share such data. Banks feel they are better off hiring a team of data scientists to work on the problem internally. These start-up ideas can die a quick death.

Despite the tremendous innovation and entrepreneurial opportunities around artificial intelligence (AI), breaking into AI can be a daunting task for entrepreneurs as they face a chicken-and-egg problem before they even begin — something existing companies are less likely to contend with. Five specific strategies can help entrepreneurs overcome this challenge and create successful AI-driven ventures.

WHAT IS THE CHICKEN-AND-EGG PROBLEM IN AI ENTREPRENEURSHIP?

Today's AI systems need to be trained on large datasets, which can pose a challenge for entrepreneurs. Established companies with a sizable customer base already have a stream of data from which they can train AI systems, build new products and enhance existing ones, generate additional data, and rinse and repeat (for example, Google Maps has over 1bn monthly active users and over 20 Petabytes of data[1]). But for entrepreneurs, the need for data poses a chicken-and-egg problem — because their company has not yet been built, they do not have data, which means they cannot create an AI product as easily.

Additionally, data is not only necessary to get started with AI, it is actually key to AI performance. Research has shown that while algorithms matter, data matters more. Among modern machine learning (ML) methods, the differences in performance between various algorithms are relatively small when compared to the performance differences between the same algorithms with more or less data.[2]

There are several strategies that can help entrepreneurs navigate this chicken-and-egg problem and access the data they need to break into the AI space (see Figure 1).

START BY OFFERING A SERVICE THAT HAS VALUE WITHOUT AI AND THAT GENERATES DATA

While data does need to come before an AI product, data does not need to come before all products. Entrepreneurs can begin by creating a service that is not AI-based, but that solves customer problems and generates data in the process. This data can later be used to train an AI system that enhances the existing service or creates a related service.

For example, Facebook did not use AI in its early days, but it still provided a social networking platform that customers wanted to join. In the process, Facebook generated a large amount of data which was in turn used to train AI systems that helped personalise the newsfeed and also made it possible to run extremely targeted ads. Despite not being an AI-driven service at the outset, Facebook has become a heavy user of AI.

Similarly, the InsurTech start-up Lemonade initially did not have data to build sophisticated AI capabilities on day one. Over time, however, Lemonade has built AI tools to create quotes, process claims and detect fraud.[3] Today, their AI system handles the 'first notice of loss' for 96 per cent of claims and manages the full claim resolution without any human involvement in a third of the cases.[4] These AI capabilities have been built using the data generated over many years of operations.

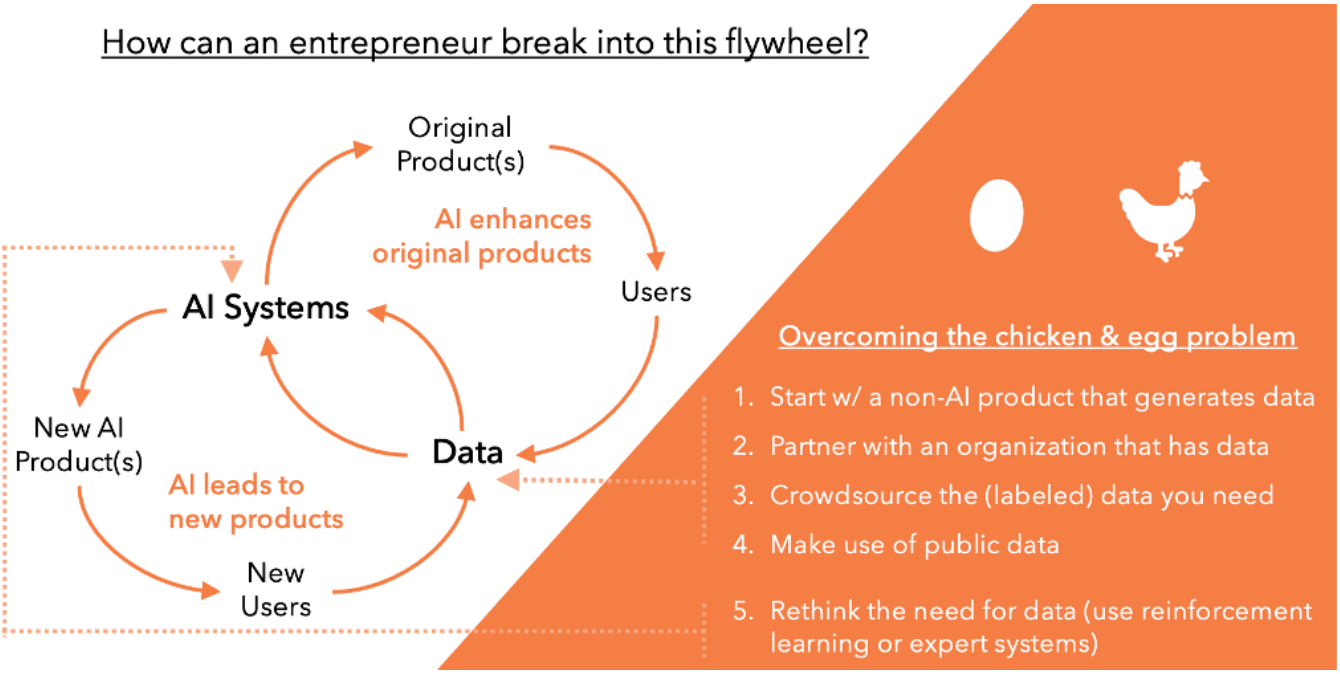
How can an entrepreneur break into this flywheel?
Original Product(s)
AI enhances original products
Users
AI Systems
Data
New AI Product(s)
AI leads to new products
New Users
Overcoming the chicken & egg problem
1. Start w/ a non-AI product that generates data
2. Partner with an organization that has data
3. Crowdsource the (labeled) data you need
4. Make use of public data
5. Rethink the need for data (use reinforcement learning or expert systems)

Figure 1: Five approaches to overcome the chicken-and-egg problem

PARTNER WITH A NON-TECH COMPANY THAT HAS A PROPRIETARY DATASET

Entrepreneurs can partner with a company or organisation that has a proprietary dataset but lacks in-house AI expertise. This approach is particularly useful in contexts where it would be very difficult to create a product that in turn generates the kind of data your AI application needs, such as medical data about patient tests and diagnoses. In this case, you could partner with a hospital or insurance company in order to obtain anonymised data.

A related point is that training data for your AI product can come from a potential customer. While this is harder in regulated industries such as healthcare and finance, customers in other industries such as manufacturing may be more open to it. All you might need to offer in return is exclusive access to the AI product for a few months or early access to future product features.

A pitfall of this approach is that potential partners may prefer working with established companies rather than smaller players who may be less known and trusted (especially in a post-GDPR and Cambridge Analytica world). So business development will be tricky, but this strategy is nonetheless feasible, especially when well-known tech companies are not already chasing after your desired partner.

Entrepreneurs who are part of a family business may already have access to a potentially large amount of data from their existing business. That is a great option too.

CROWDSOURCE THE (LABELLED) DATA YOU NEED

Depending on the kind of data needed, entrepreneurs can obtain data through crowdsourcing. When data is available but is not well labelled (eg images on the Internet), crowdsourcing can be a particularly well-suited method for obtaining this data, as labelling is a task that lends itself well to being completed quickly by a large number of individuals on crowdsourcing platforms.

Platforms such as Amazon Mechanical Turk and Scale.ai are frequently used to help generate labelled training data.

For example, consider Google's use of Captchas. While they serve an important security purpose, Google simultaneously uses them as a crowdsourced image labelling system. Every day 'millions of users are part of the Google analytics pre-processing team which are validating machine learning algorithms for free'.[5]

Some products have workflows that allow customers to help label new data in the course of using the product. In fact, the entire subfield of active learning is focused on how to interactively query users to better label new data points. For example, consider a cyber security product that generates alerts about risks and a workflow in which an ops engineer resolves those alerts, thereby generating new labelled data. Similarly, product recommendation services such as Pandora use upvotes and downvotes to validate recommendation accuracy. In both these cases, you can start with a minimum viable product (MVP) that continually improves over time as customers provide feedback.

MAKE USE OF PUBLIC DATA

Before you conclude that the data you need is not available, look harder. There is more publicly available data than you might imagine; there are even data marketplaces emerging. While publicly available data (and therefore the resulting product) might be less defensible, you can build defensibility through other service/product innovations, such as creating an exceptional user experience or combining offline and digital data at scale, as Zillow does (the company uses offline public municipal data at scale as part of their innovative online real estate application[6]). One could also combine publicly available data with some proprietary data, which could be generated over time or obtained through partnerships, crowdsourcing, etc.

The Canadian company BlueDot uses a variety of data sources, including publicly available data, in order to detect outbreaks of emerging diseases before they are officially reported as well as predict where an outbreak will spread to next. BlueDot uses 'statements from official public health organizations, digital media, global airline ticketing data, livestock health reports, and population demographics', among other data sources. The company detected the COVID-19 outbreak on 30th December, 2019, nine days before the WHO reported on it.[7]

In addition to data, there is also the option of using publicly available pre-trained ML models that can be customised to your needs with transfer learning. Transfer learning involves the use of a model developed for one task as the starting point for another task. For example, a deep learning model trained on millions of images can be the starting point for a more domain-specific image recognition model (eg to identify flowers). This approach is commonly used for image recognition and natural language processing and can significantly reduce the need for new data.

RETHINK THE NEED FOR DATA

It is true that most of the practical AI in the business world is based on ML, and most of that is supervised ML (which requires large labelled training datasets). But many problems can be solved with other AI techniques that are not reliant on data, such as reinforcement learning or expert systems.

Reinforcement learning is an ML approach in which algorithms learn by testing various actions or strategies and observing the rewards from these actions. Essentially, reinforcement learning uses experimentation to compensate for a lack of labelled training data. The original iteration of Google's Go playing software, AlphaGo, was trained on a large training dataset, but the next iteration, AlphaZero, was based on reinforcement learning and had zero training data. Yet AlphaZero beat AlphaGo (which itself beat Lee Sedol, Go's world champion).

Reinforcement learning is widely used in online personalisation. Online companies frequently test and evaluate multiple website designs, product descriptions, product images and pricing. Reinforcement learning algorithms explore new design and marketing choices and rapidly learn how to personalise user experience based on their responses.

Another approach is to use expert systems, which are simple rule-based systems that often codify rules that experts use routinely. While expert systems rarely beat well-trained ML systems for complex tasks such as medical diagnosis or image recognition, they can help break the chicken-and-egg problem and help you get started. For example, the virtual healthcare company Curai used knowledge from expert systems to create clinical vignettes, and then used these vignettes as training data for ML models (alongside data from electronic health records and other sources).[8]

To be clear, not every intelligence problem can be cast as a reinforcement learning problem or tackled through an expert systems approach, but these are worth considering when the lack of training data has halted the development of an interesting ML product.

Entrepreneurs are most likely to develop a consistent stream of proprietary data if they start by offering a service that has value without AI and that generates data, and then use this to train an AI system. This strategy does require time and may not be the best fit for all situations. Depending on the nature of the start-up and the kind of data that is needed, it may work better to partner with a non-tech company that has a proprietary dataset, crowdsource (labelled) data, or make use of public data. Alternatively, entrepreneurs can rethink the need for data entirely and consider taking a reinforcement learning or expert systems approach.

ACKNOWLEDGMENTS

Thanks to Jocelyn Goldfein of Zetta Venture Partners for many helpful suggestions and Wharton student Anirudh Viswanathan for feedback. Thanks to Quentin Demersay for the suggestion to add publicly available ML models (and transfer learning) to the list of strategies. An earlier version of this article was posted on Towards Data Science.[9] All errors are my own.

References

1. Nahar, A. (April 2017), 'Google Maps – the most expansive data machine', Digital Innovation and Transformation, available at https://digital.hbs.edu/platform-digit/submission/google-maps-the-most-expansive-data-machine/ (accessed 14th June, 2021).
2. Banko, M. and Brill, E. (2001), 'Scaling to very, very large corpora for natural language disambiguation', Microsoft Research, available at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/acl2001.pdf (accessed 14th June, 2021).
3. Wininger, S. (January 2020), 'The sixth sense. Lemonade's 2019 product in review: Blond wigs, weird bots, and ad fails', Lemonade, available at https://www.lemonade.com/blog/the-sixth-sense/ (accessed 14th June, 2021).
4. Lemonade, Inc. (June 2020), 'Form S-1', available at https://www.sec.gov/Archives/edgar/data/1691421/000104746920003846/a2241899zs-1a.htm. (accessed 14th June, 2021).
5. Boer, M. (July 2020), 'AI-labeling crowd-sourcing platforms', Medium, available at https://medium.com/swlh/ai-labeling-crowdsourcing-platforms-630adbc79c40 (accessed 14th June, 2021).
6. Zillow (n.d.), 'Where does Zillow get information about my property?', Zillow Help Center, available at https://zillow.zendesk.com/hc/en-us/articles/213218507-Where-does-Zillow-get-information-about-my-property- (accessed 14th June, 2021).
7. Stieg, C. (March 2020), 'How this Canadian startup spotted coronavirus before everyone else knew about it', CNBC, available at https://www.cnbc.com/2020/03/03/bluedot-used-artificial-intelligence-to-predict-coronavirus-spread.html (accessed 14th June, 2021).
8. Kannan, A. (April 2019), 'The science of assisting medical diagnosis: From expert systems to machine-learned models', Curai Tech Blog, Medium, available at https://medium.com/curai-tech/the-science-of-assisting-medical-diagnosis-from-expert-systems-to-machine-learned-models-cc2ef0b03098 (accessed 14th June, 2021).
9. An earlier version of this paper was published at https://towardsdatascience.com/how-to-kickstart-an-ai-venture-without-proprietary-data-13d1502051f2 (accessed 14th June, 2021).